PATHWAYS TO INSIGHT

INTEGRATED BIOLOGY AT AGILENT

The Measure of Confidence

Agilent Technologies

"Biological research is expanding enormously in its ability

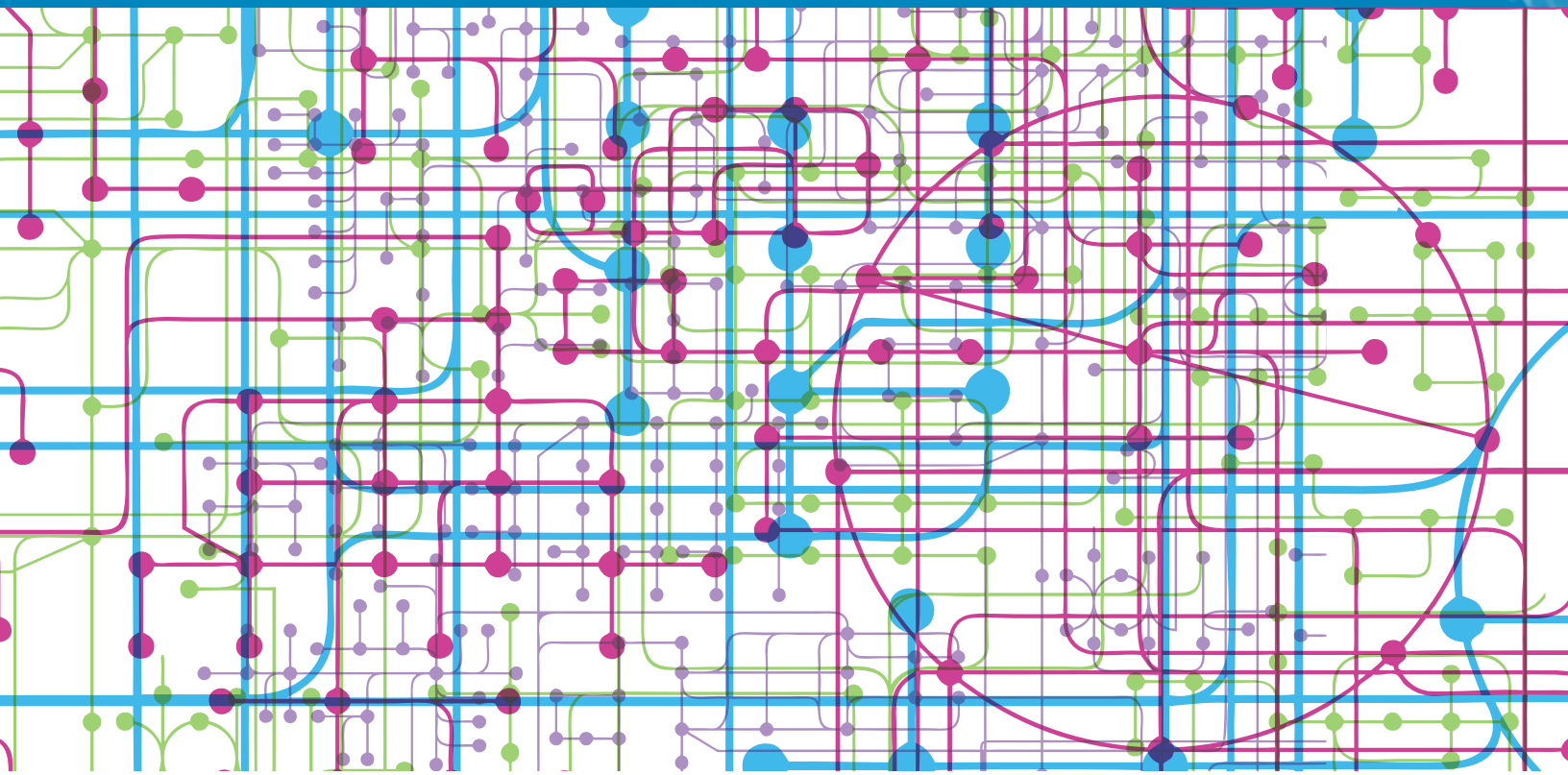
diverse and complementary data types and to inject prior

TREY IDEKER, PHD, DEPARTMENTS OF MEDICINE AND

to decipher complex systems. This ability derives from the expanded power to incorporate understanding of biological principles."[1]

BIOENGINEERING, INSTITUTE FOR GENOMIC MEDICINE, UNIVERSITY OF CALIFORNIA, SAN DIEGO, LA JOLLA, CA, USA

RISE ABOVE THE NOISE



Answers Emerge Through Integrated Analyses

The complexity of biology continues to challenge researchers' ability to find answers through omics approaches.[1] While genomics, transcriptomics, proteomics, and metabolomics are in wide use in both industry and academia, these experiments—performed alone—often lack the statistical power to uncover meaningful correlations amid the high levels of noise omics experiments typically generate. For example, "the millions of single-nucleotide variants (SNVs) found in a typical genome-wide association study[2] make it extremely difficult to identify which particular SNVs are the true causes of disease."[1]

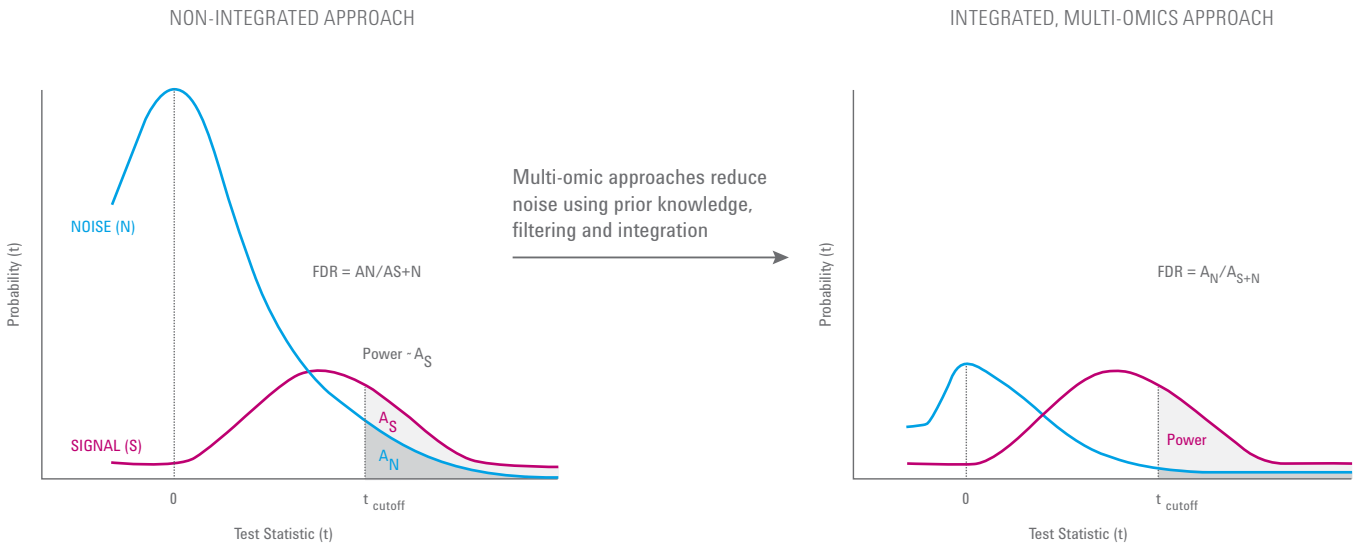
In a recent review discussing how to overcome this challenge, Ideker, *et al*,1 identify two ways to increase signal-to-noise in omics studies: (1) incorporate prior knowledge about the biological system such as known interactions and pathways, and (2) integrate "complementary layers" of data from the genome, transcriptome, proteome, metabolome, and interactome. In their review, the authors highlight four studies that demonstrate how these types of integrated analyses are transforming our understanding of biology and leading to groundbreaking new insights.

Agilent solutions for integrated biology

At Agilent, we agree that integrated, multi-omics approaches are a powerful and important tool for biological researchers. We continue to support these types of approaches through our Integrated Biology solutions, our Integrated Biology Grant Program, and our commitment as a partner to open-source projects like Cytoscape (www.cytoscape.org).

With analytical products across the four major omics—genomics, transcriptomics, proteomics, and metabolomics—we are uniquely positioned among life science companies to enable integrated multi-omics approaches. And the latest version of our GeneSpring bioinformatics software suite makes possible truly integrated, pathway-level analysis of primary data from any of our omics platforms, while also enabling incorporation of prior knowledge—existing datasets, pathway maps, and interaction maps—for greater analytical power in multi-omics experiments.

Here we discuss Agilent's solutions for Integrated Biology and illustrate how leading researchers are using our technology to answer questions in cancer biology, infectious diseases, and systems toxicology. Of course, integrated approaches are not limited to these fields but can be expanded to study the full range of biological systems including plants, humans, microbes and model and non-model organisms. To learn more about Agilent's vision of Integrated Biology, continue reading, visit us on the web, or contact an Agilent representative.



The statistical power of omics experiments can be enhanced through bioinformatics methods that decrease noise through the use of (1) complementary datasets, and (2) incorporation of prior knowledge about the system (e.g., aggregating measurements from entities that belong to the same pathway). This results in an effective decrease in the False Discovery Rate (FDR), at a given t statistic cutoff, or in the ability to relax such cutoff while maintaining the same FDR.

EMBRACE INTEGRATED APPROACHES

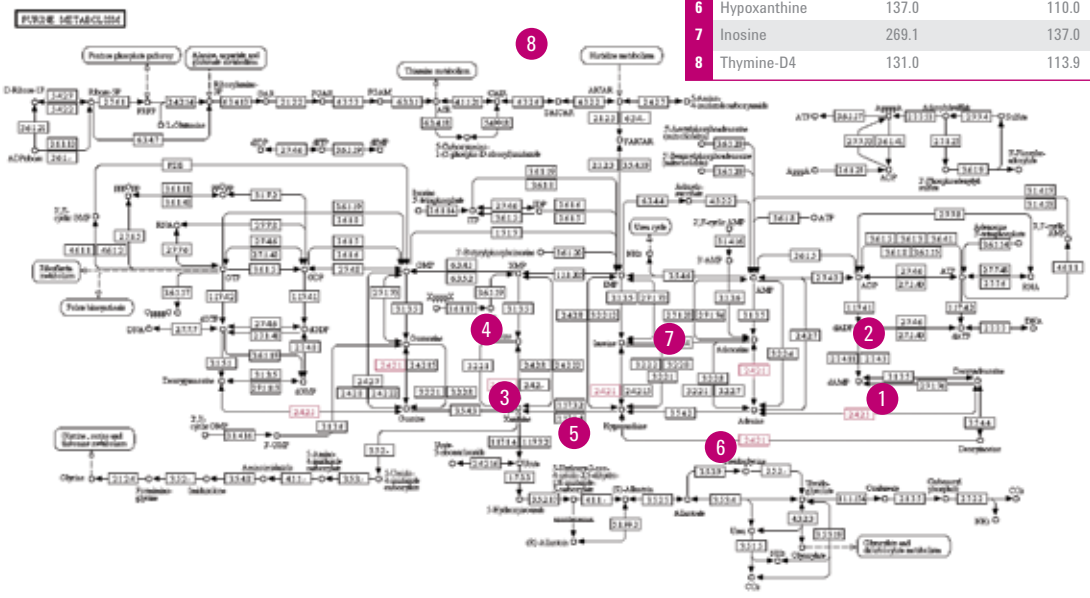
A Faster, More Reliable Journey From Discovery to Pathway-Driven Validation

Omics experiments generate large numbers of measurements, and it is often difficult to determine which events correlate with a specific disease or phenotype. Building on the existing knowledge about a system can help identify the true correlates—the signal in the noise—and lay out the path for the next round of experiments that validate the initial findings. The work of Arun Sreekumar, PhD, now at Baylor College of Medicine, provides an elegant example of this pathway-driven process as he searched for biomarkers for pancreatic cancer (pancreatic duct adenocarcinoma, PDAC).[4]

The group's work consisted of three phases:

- a hypothesis-free search for protein signatures that correlate with disease
- an analysis and initial validation of the protein signatures via secondary methods, and
- validation experiments using an orthogonal targeted metabolomics approach, designed using existing knowledge about the biological pathways that the suspected proteins are involved in.

	ANALYTE	PARENT IONS (m/z)	PRODUCT IONS (m/z)
1	Adenine	136.6	119.0
2	Adenosine	268.1	136.0
3	Guanine	152.0	135.0
4	Guanosine	284.1	152.0
5	Xanthine	153.0	110.0
6	Hypoxanthine	137.0	110.0
7	Inosine	269.1	137.0
8	Thymine-D4	131.0	113.9



Testing their hypothesis that nucleoside phosphorylase (NP) is involved in PDAC, Sreekumar's group identified seven metabolites that are either NP substrates or products (1-7), and one distally upstream metabolite as a control (8).[4]

Phase 1: Proteomic profiling for discovery

Looking at pancreatic juice samples from benign (n=7), PDAC (n=16), and pre-cancerous tissues (carcinoma *in situ* (cis), n=2), Sreekumar's group looked for class-specific protein expression signatures—the set of proteins whose differential expression correlates with one of the three groups. After normalizing and analyzing the MS data using a two-sided t-test coupled to a false discovery rate (FDR) correction, the group identified 56 (out of 431) proteins that are significantly upregulated in PDAC ($p \leq 0.05$, FDR≤12%) when compared to benign samples.

Phase 2: Validating class-specific protein signatures

To ensure that the results from the proteomics experiment reflected true biological, disease-specific differences, Sreekumar's group assessed expression of a subset of their class-specific proteins using tissue microarrays and immunoblot analysis.

Phase 3: Protein biomarker verification through pathway-driven targeted metabolomics

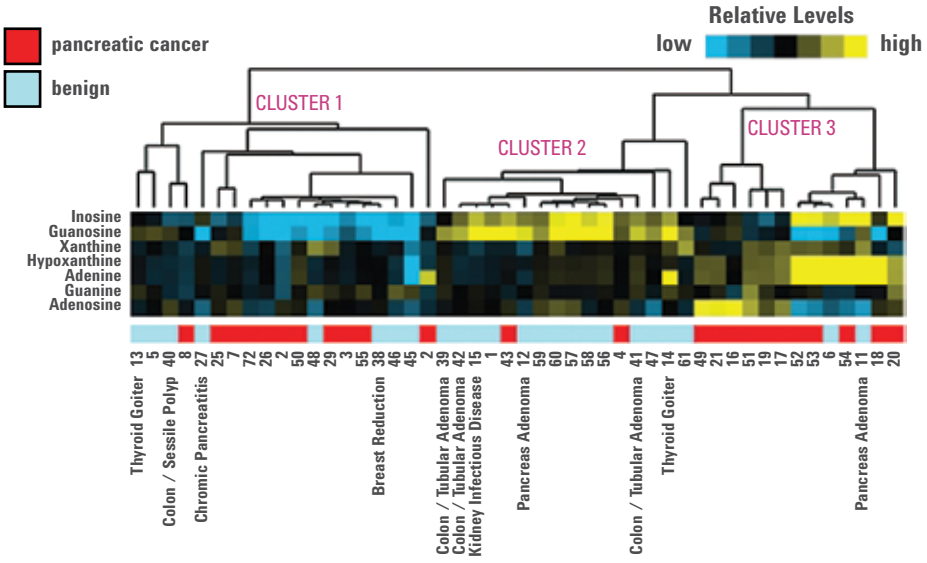
After validating their protein expression data, Sreekumar's group selected a set of proteins for further validation, including purine nucleoside phosphorylase (NP). NP, a component of the purine salvage pathway, is extensively used by tumors to replenish their nucleotide pools. Sreekumar's group examined the purine metabolic pathway and identified seven metabolites that are either NP substrates or products, and one metabolite distally upstream of NP. Using a targeted LC/MS/MS metabolomics approach, they measured the levels of these metabolites and were able to construct a biomarker panel to differentiate between the two sample classes—benign versus PDAC—with high statistical confidence ($p = 0.00025$, one-sided Fisher's exact test).

The work of Arun Sreekumar illustrates one way to incorporate existing knowledge to identify meaningful correlations in omics studies. The key to singling out NP as a biomarker lay in the use of existing knowledge of NP's function in specific cellular pathways. By examining the pathways NP is involved in, Sreekumar's group was able to quickly formulate a round of validation experiments using a targeted, pathway-driven metabolomics approach.

Simpify with Agilent GeneSpring

Agilent's Integrated Biology software (GeneSpring-IB) simplifies this type of approach through the use of the Pathway Architect (PA) module. GeneSpring-IB jointly processes the raw data from genomics, transcriptomics, proteomics, and metabolomics studies—two omics at a time—and maps differentially expressed entities onto their corresponding pathway. This information gets passed to the PA, where users can intuitively explore, verify, and quickly find the biological pathways overrepresented in the joint analysis of their multi-omics datasets.

Integrating pre-existing data and using multi-omics approaches speeds the path from discovery and insight to validation and application.



Hierarchical clustering of the NP metabolite panel allowed successful differentiation of benign and PDAC samples.4

INTEGRATE COMPLEMENTARY OMICS APPROACHES

More Efficient Discovery

There are times when a single omics approach does not provide enough signal for statistically valid identification of a discrete set of target entities. However, integration of data from multiple omics can, in some cases, provide enough constraints to greatly reduce the FDR. The PA module of GeneSpring allows either single omics analysis or joint analysis of multiple omics, increasing your ability to find reliable answers quickly and accurately.

Joint analysis for more reliable results—a case study

Akhilesh Pandey, MD, PhD, Johns Hopkins University, and his group recently combined transcriptomic data with proteomic profiling to uncover the mechanism of action of sulforaphane (SFN), a potential chemopreventative for breast cancer, and to identify potential pharmacodynamic biomarkers.[5]

Kelch-like ECH-Associated Protein 1 (KEAP1) is a known target of SFN. Binding of SFN to KEAP1 inactivates KEAP1's inhibitory function, allowing induction of detoxification and antioxidant enzymes. Pandey's group examined both gene expression and protein expression in cells treated with SFN or in KEAP1 knockdown cells. Gene expression analysis resulted in 6,378 transcripts regulated by SFN and 1,710 transcripts regulated by KEAP1 knockdown, with 879 transcripts overlapping.

From the proteomics experiments, SFN treatment resulted in detection of 96 upregulated proteins and 26 downregulated, compared to 50 upregulated and 76 downregulated in the KEAP1 knockdown. The overlap was 29 differentially regulated proteins.

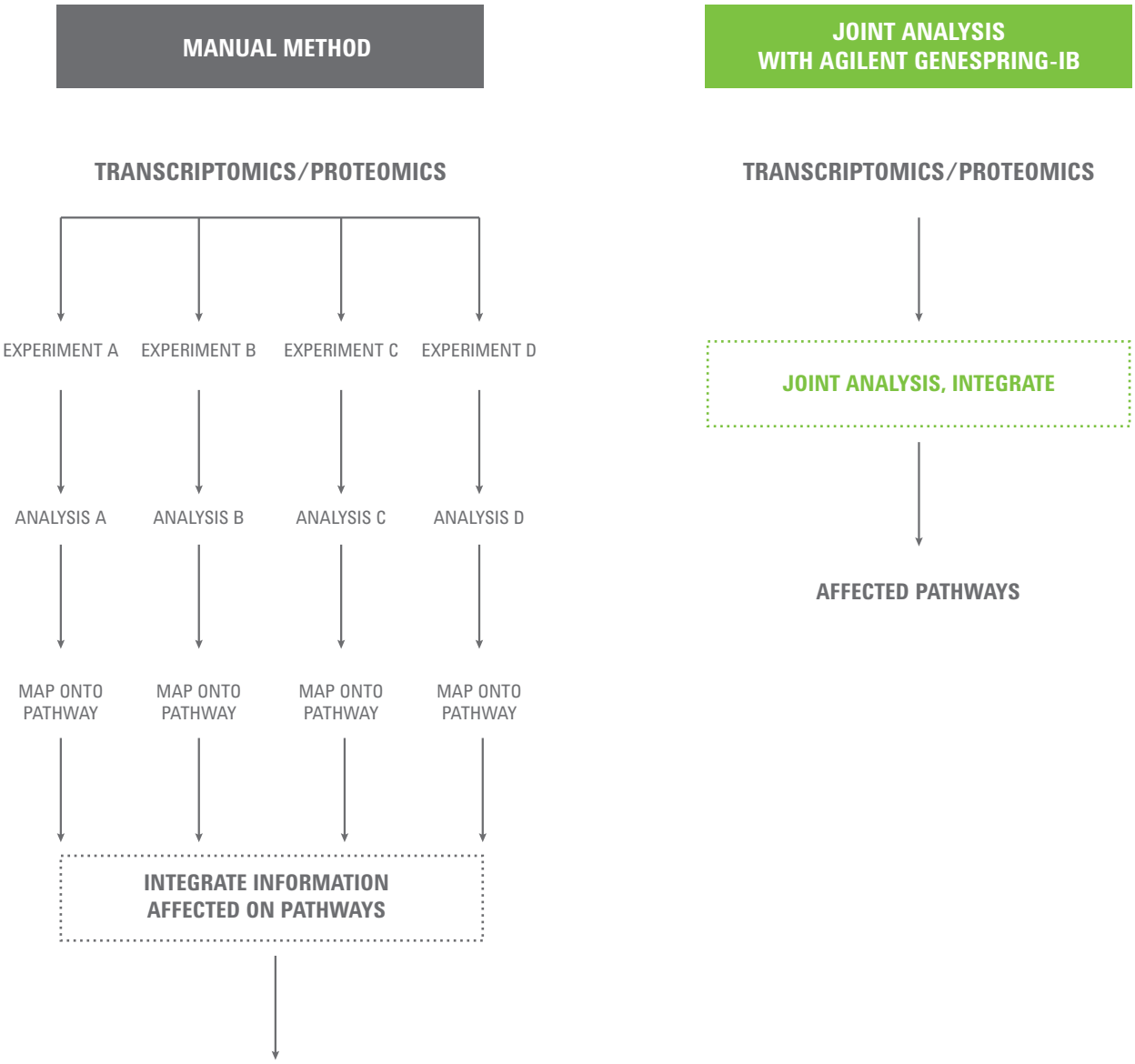
When combining all four experiment types, Pandey's group was able to identify commonly affected pathways—xenobiotic metabolism and antioxidants, glutathione metabolism, carbohydrate metabolism, and NADH/NADPH regeneration—and a single family of proteins, the aldo–keto reductase family members—AKR1B10, AKR1C1, AKR1C2 and AKR1C3, as well as NQO1 and ALDH3A1—that had both genes and proteins upregulated in both SFN-treated and KEAP1 knockdown cells.

By assessing two different conditions using two different omics techniques, Pandey's group was able to reduce FDR and pare down the set of potential biomarkers from almost a thousand to only a handful of highly significant targets.

Pandey's group had to manually analyze and integrate all of their data. With GeneSpring-IB, joint analysis is easier and faster, enabling statistical, significant insights with less effort.

	TRANSCRIPTOMICS	PROTEOMICS
SFN treatment	6, 378	122
KEAP1 knockdown	1,710	126

Pandey's group integrated data from two different omics technologies and two different conditions to reduce FDR and identify the most promising biomarker candidates.[5]



Agilent GeneSpring-IB makes it easier to perform joint analysis of omics data.

Researchers in Infectious Diseases and Toxicology Embark on Multi-omics Approaches

Integrated, multi-omics approaches are powerful tools for understanding complex biological systems, helping researchers find meaningful signals amid the background noise from cellular processes. As the tools for integrating omics become easier to use, researchers from every field are increasingly turning to integrated studies to find answers.

Systems toxicology

We're also helping Russell Thomas, PhD, The Hamner Institutes for Health Sciences, add proteomics and metabolomics approaches to his transcriptomics studies of pathways involved in toxicology.[6] Both European and US governments have recently developed initiatives to drastically increase the number of compounds routinely screened for toxicity.[7] These policies challenge the throughput of current toxicological testing methods and require the use of a prohibitive number of animals. Toxicologists like Russell Thomas and Thomas Hartung, MD, PhD, Johns Hopkins Bloomberg School of Public Health, are developing integrated testing and multi-omics approaches to identify biomarkers for toxicity that can be used in *in vitro*, cell-based tests. Not only will these tests allow high-throughput screening of compounds for toxic effects, they are expected to provide more accurate assessment of toxicity in humans than current animal-based approaches.

"...[omics] technologies not only allow researchers to 'fish' for new biological markers of specific toxic effects but also increasingly allow the deduction of patterns (or signatures) that are characteristic of certain toxic effects. By also harnessing advances in bioinformatics and in silico modelling, this information can be mined and then integrated with knowledge from other areas of the life sciences."[7]

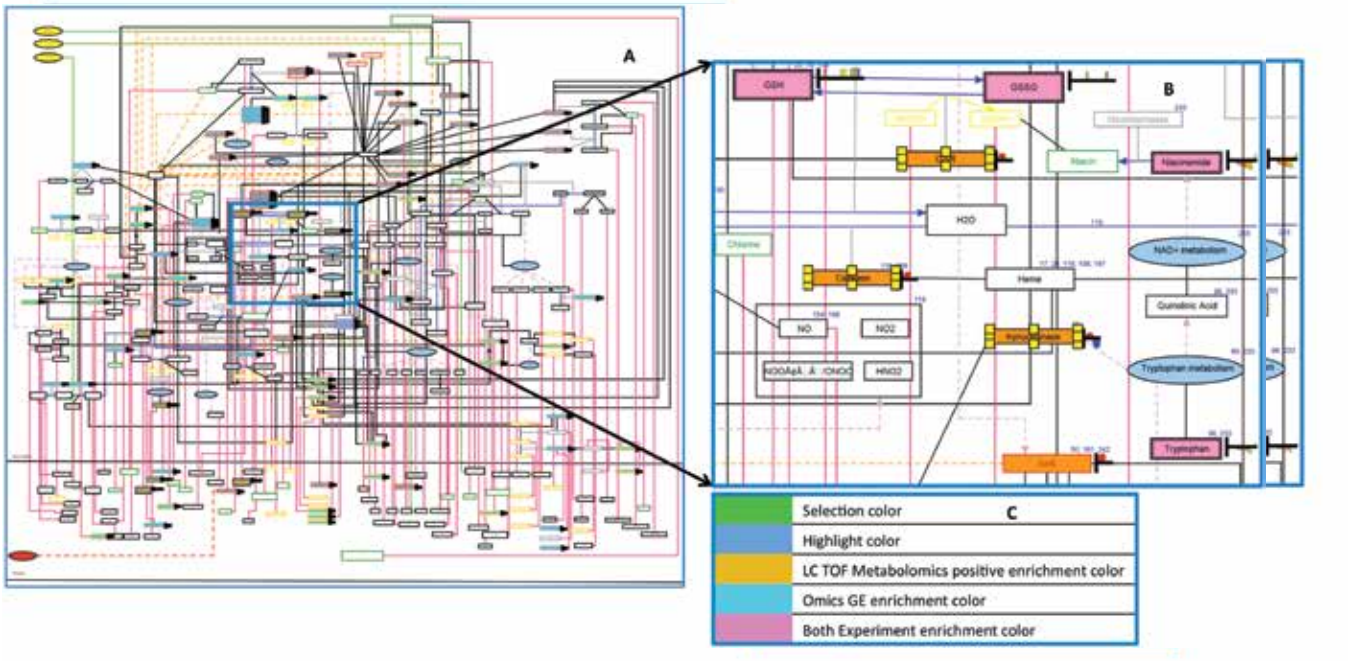
THOMAS HARTUNG, MD, PHD, BLOOMBERG SCHOOL OF PUBLIC HEALTH, JOHNS HOPKINS UNIVERSITY, BALTIMORE, MARYLAND, USA

Infectious disease research

Agilent is supporting noted tuberculosis researcher Kyu Rhee, MD, PhD, Weill Cornell Medical College, as he adds transcriptomics analysis to his metabolomics approaches.[8] Rhee and colleagues have already used both targeted and discovery metabolomics approaches to uncover novel ways that M. *tuberculosis* metabolizes different carbon sources—unlike other bacteria, M. *tuberculosis* can utilize multiple carbon sources at the same time, suggesting that drugs that target a single metabolic pathway will not be completely effective. By combining transcriptomics approaches with metabolomics, Rhee and colleagues hope to learn more about the biology of this re-emerging infectious disease, informing the development of new, more effective therapeutics.

Implementing multi-omics studies

With our expert application scientists, product lines in all the major omics, and GeneSpring-IB software to integrate and analyze multi-omics data, Agilent can help researchers expand into multi-omics projects.



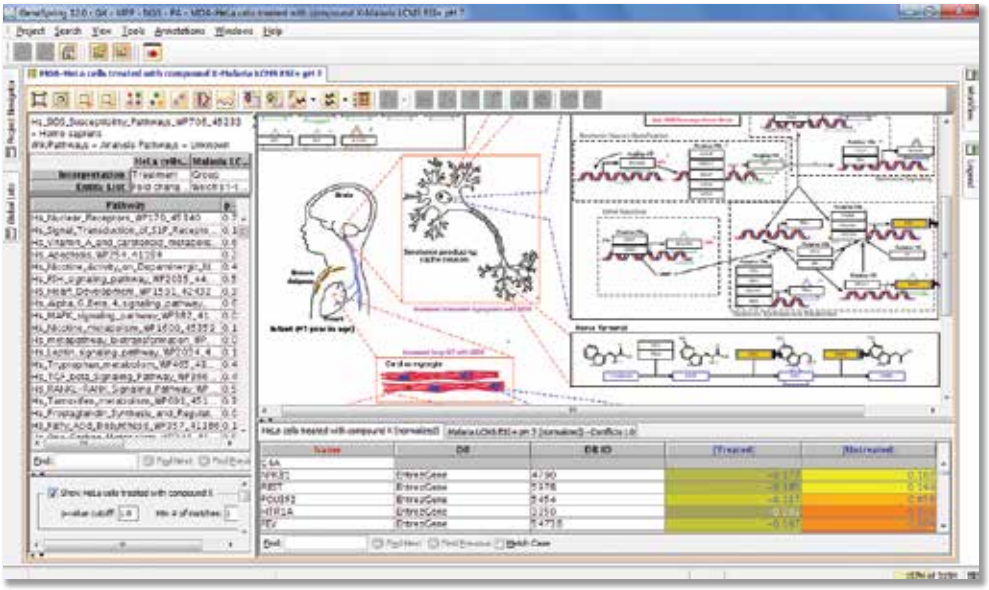
Combined analysis of genomics and metabolomics data using Agilent GeneSpring 12.0 – MPP. (A) Selenium pathway and (B) expanded detail highlighting some of the differentially expressed metabolites and genes (yellow squares) in the pathway. (C) Legend showing the color settings for the pathway view and the heatstrip.[9]

FIND DEEPER INSIGHT

Agilent GeneSpring's Pathway Architect Module Enables Integrated Biology

At the heart of Agilent's Integrated Biology offering is the GeneSpring Bioinformatics Software Suite. Through fully integrated modules for analyzing genomics, transcriptomics, proteomics, and metabolomics, GeneSpring provides intuitive, easy-to-use statistical analyses and biological contextualizationof single omics studies. Combined with extensive visualization capabilities and robust data management architecture, GeneSpring acts as a central analytical resource. It's also easily customizable through R and Jython scripting. Cited over 13,000 times, researchers around the world have been using GeneSpring to gain deeper insight into their data.

Addition of the Pathway Architect (PA) module to GeneSpring allows multi-omics integration—genomics, transcriptomics, proteomics, and metabolomics data can be jointly analyzed to increase statistical power and to facilitate the move from hypothesis-free to hypothesis-driven studies, as in Arun Sreekumar's (p. 6) and Akhilesh Pandey's (p. 8) work. The PA module builds upon the biological analysis from any of the omics-level modules, and leverages the expertise of the larger biological community through seamless incorporation of the WikiPathways resource. Primary data analysis from single or multi-omics experiments can be used to identify pathways involved in the process under study—such as cellular state, disease progression, or toxicity—and because the PA module is fully integrated, these studies easily transition from discovery to design of the validation round of experiments.

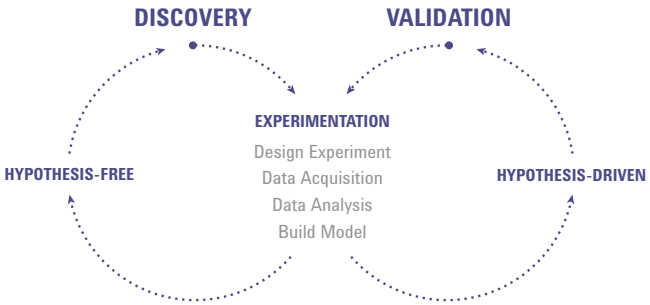


Agilent GeneSpring is a powerful, widely cited tool for omics data analysis.

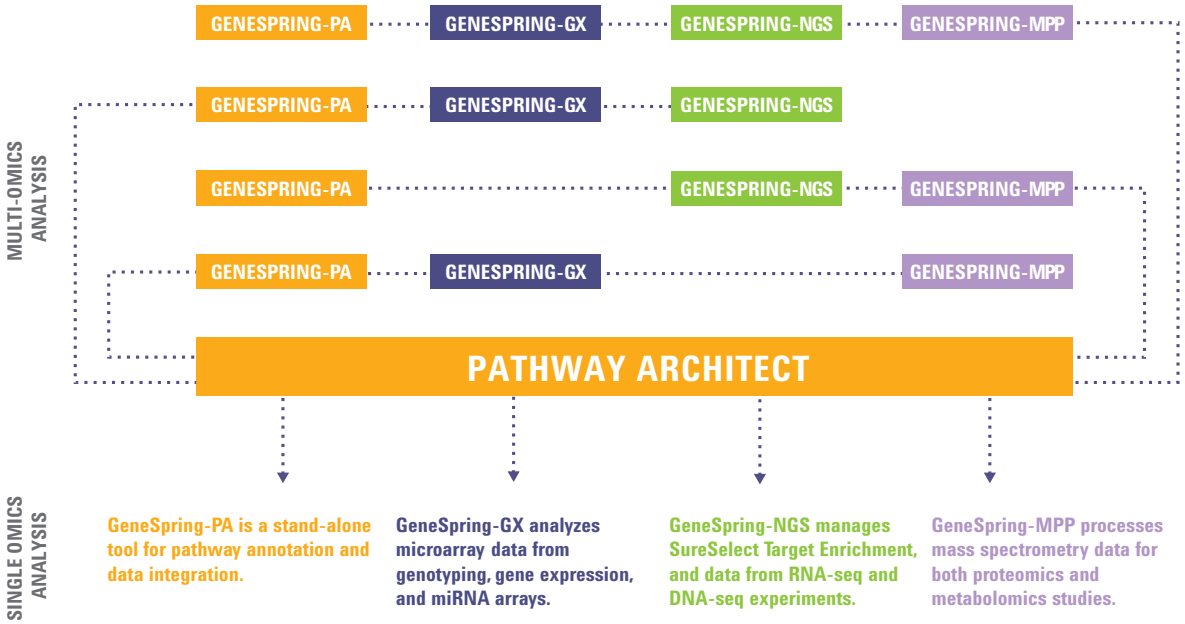
The flexibility of the PA module allows researchers to use multiple methods for identifying relevant biological pathways:

- Use the preloaded WikiPathways, a resource that is continuously curated by the biological research community
- Use built-in natural language processing-based (NLP) algorithms and medical subject heading (MeSH) terms across PubMed, MEDLINE, and proprietary sources to identify new networks
- Upload custom pathway information from nearly any source
- Build pathways from primary data using neural network analysis
- Automatically translate pathway identifiers between measurement platforms, annotation sources, and different organisms
- Transfer data between GeneSpring and Ingenuity Pathways Analysis (IPA)
- Export data from GeneSpring to Thomson Reuters' GeneGo MetaCore analysis suite

Agilent supports your multi-omics Integrated Biology projects at every step, from data acquisition to data analysis and visualization. Our powerful GeneSpring-IB software platform, configurable with the NGS, GX, MPP, and PA modules, performs statistical and pathway analyses, and, unique among omics software tools, can jointly analyze your raw omics data, increasing signal-to-noise. Agilent's solutions for Integrated Biology take you from question to insight quickly and confidently.



Agilent supports and integrates single and multi-omics experiments, simplifying the researcher's cyclical journey between hypothesis-free discovery and pathway/hypothesis-driven validation.



The Agilent GeneSpring suite of Bioinformatics Software tools is highly modular and configurable. Combine different single omics modules for integrated analysis.

GLOSSARY OF OMICS

Genomics—Obtaining DNA sequence information and assessing genetic variation at the DNA sequence and genomic architecture levels

Agilent offers a wide range of systems, reagents, and consumables, including CGH and CNV microarray platforms, PCR and qPCR systems, and SureSelect Target Enrichment reagents for next-gen sequencing.

Transcriptomics—Measuring gene expression levels and uncovering differential expression of mRNA and miRNA

Agilent offers a number of systems, reagents, and consumables, including microarrays and microarray platforms.

Proteomics—Qualitative and quantitative determination of protein expression and modification, usually using mass spectrometry

Agilent offers a portfolio of LC/MS instruments for both discovery and targeted proteomics.

Metabolomics—Following changes in the identity and levels of metabolites in a sample

Agilent offers a portfolio of LC/MS and GC/MS instruments for both discovery and targeted metabolomics.

AGILENT'S COMMITMENT TO INTEGRATED BIOLOGY

Forging a Path Into the Future

Omics technologies have matured and researchers are now beginning to extract even more information from omics data through integrated approaches. At Agilent, we are not only committed to helping researchers begin their own journeys of integrated, multi-omic discovery, we are committed to moving the field of Integrated Biology forward.

We are a major supporter of Cytoscape, an open-source platform for visualization and analysis of complex networks. Through funding and collaboration, we help ensure that the entire research community retains free and open access to this invaluable research tool.

We have also awarded our first sets of grants to fund development of open-source software tools that enable integrated approaches. Congratulations to Peter J. Park, Harvard University, and Michael J. MacCoss, University of Washington. Please visit our eMerging Insights Grants pages for more information on Park and MacCoss, and updates on new grant programs.[9]

Integrated approaches to biology are powerful ways to address important questions in basic and applied biological research. They increase the statistical power of your data—providing higher signal-to-noise ratios—and let you see farther by standing on the shoulders of the researchers who've gone before. To learn more about Integrated Biology at Agilent visit agilent.com,[4] or speak with an Agilent representative.

References

1. Ideker, T., Dutkowski, J. & Hood, L. Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell* **144**: 860-863 (2011).

2. Hudson, T.J. *et al*. International network of cancer genome projects. *Nature* **464**: 993-998 (2010).

3. Integrated Biology at Agilent <http://www.agilent.com/lifesciences/biology>

4. Vareed, S.K. *et al*. Metabolites of purine nucleoside phosphorylase (NP) in serum have the potential to delineate pancreatic adenocarcinoma. *PLoS ONE* **6**: e17177 (2011).

5. Agyeman, A.S. ***et al***. Transcriptomic and proteomic profiling of KEAP1 disrupted and sulforaphane-treated human breast epithelial cells reveals common expression profiles. *Breast Cancer Res. Treat*. (2011). **132** (1): 175–187 (2012).

6. Thomas, R.S. *et al*. Application of transcriptional benchmark dose values in quantitative cancer and non-cancer risk assessment. *Toxicol. Sci.* **120**: 194-205 (2011).

7. Hartung, T. Toxicology for the twenty-first century. *Nature* **460**: 208-212 (2009).

8. de Carvalho, L.P.S. *et al*. Metabolomics of Mycobacterium tuberculosis reveals compartmentalized co-catabolism of carbon substrates. *Chem. Biol.* **17**: 1122-1131 (2010).

9. eMerging Insights Grant Program <http://www.agilent.com/lifesciences/emerginginsights>

Learn more
www.agilent.com/lifesciences/biology

Buy online
www.agilent.com/chem/store

Find an Agilent customer center in your country
www.agilent.com/chem/contact

U.S. and Canada
1-800-227-9770
agilent_inquiries@agilent.com

Europe
info_agilent@agilent.com

Asia Pacific
inquiry_lsca@agilent.com

For Research Use Only. Not for use in diagnostic procedures.
This information is subject to change without notice.

Printed in the USA, April 22, 2015
5991-0222EN

Agilent Technologies